UNIT 6 · BLOCK 2

Correlation

Do two things rise and fall together — and how tightly?

Why this matters

So far we've described **one** variable at a time. But the most interesting questions in psychology are about **pairs**: does more sleep go with less stress? Does more phone time go with more loneliness? **Correlation** is the single number that answers "do these two move together — and how tightly?" It's the workhorse of psychological research — and the most *misread* number in all of science, because a relationship between two things is not the same as one causing the other. One Google Sheets function does the arithmetic; this unit teaches you what the number actually means.

By the end you can...

- Say what a **correlation** measures — the **direction** and **strength** of a straight-line relationship.
- Read the coefficient ***r*** — its **sign** (direction) and its **size** (strength), always between **−1 and +1**.
- Read a **scatterplot** at a glance: uphill, downhill, or shapeless blob.
- Explain why **correlation does not equal causation** — third variables and directionality.

Spot the two traps that fool *r*: **restricted range** and **outliers**.

1. The big idea — two questions at once

A **correlation** describes how two **quantitative** variables move together. It compresses a whole cloud of data points into a single number — the **correlation coefficient**, written ***r*** — that answers two separate questions at the same time:

Question	Answered by	In plain English
Which **direction**?	the **sign** of *r* (+ or −)	up together, or opposite ways?
How **strong**?	the **size** of *r* (how close to ±1)	tight line, or scattered cloud?

The version you'll use almost always is **Pearson's *r*** (full name: the *Pearson product-moment correlation coefficient*). It's built to detect **straight-line** patterns — a steady "as one goes up, the other tends to go up (or down)." Keep that word *straight-line* in mind: it's the source of *r*'s biggest blind spot, which we'll meet later.

Read it as two answers, never one

Whenever you see an *r*, split it in two on the spot: the **sign** tells you the *direction*; the **size** tells you the *strength*. They are independent questions, and *r* packs both into one value.

2. The coefficient *r* — direction (the sign)

r is a number that *always* lands between **−1 and +1**. It can never be 1.4, never −2 — if your arithmetic ever lands outside that range, you made a mistake. The **sign** tells you the direction of the relationship.

- **Positive (+):** the two go **up together**. More study hours, higher GPA. As x rises, y tends to rise.
- **Negative (−):** as one goes **up**, the other goes **down**. More sleep, less stress. As x rises, y tends to fall.
- **Zero (≈ 0):** no straight-line link at all. Knowing x tells you nothing about y.

A plus isn't "good"

A plus or a minus isn't "good" or "bad" — it's just **which way** the relationship points. A negative *r* is a real, strong relationship, not a worse one. "Sleep goes with *less* stress" (*r* negative) is every bit as strong and real as "study goes with *more* GPA" (*r* positive).

CHECK **A classmate reports a correlation of *r* = 1.4 between sleep and mood. What's wrong?**

Impossible. Pearson's *r* is mathematically bounded between −1 and +1. Any value outside that range means an error in the calculation, not an unusually strong relationship. The strongest a positive correlation can ever be is exactly +1.

3. The coefficient *r* — strength (the size)

The **size** of *r* — its distance from zero, ignoring the sign for a moment — tells you how **tight** the relationship is. Close to **±1**, the dots hug a straight line: strong and dependable, so knowing one score lets you predict the other well. Close to **0**, the dots are a scattered mess: weak or no relationship, and one score tells you almost nothing about the other.

There's no law of nature that sets the cutoffs, but psychologists lean on a rough convention (often credited to Cohen) for talking about **effect size**. Treat it as a *guide*, not gospel — and read it on the **absolute value** |*r*|, since direction doesn't change strength:

|*r*| roughly…	Strength (rough guide)	What the cloud looks like
.00 – .10	none / trivial	shapeless blob, no tilt
~.10	weak (small)	faint tilt, very fuzzy
~.30	moderate (medium)	clear tilt, still scattered
~.50 and up	strong (large)	tight band around a line

So ***r* = +.52** is a strong, positive relationship. ***r* = −.46** is a strong, negative one — the minus is just direction. And ***r* = −.02** is basically **nothing**: knowing x leaves you no better off at guessing y. In psychology, where humans are messy and many causes pile up, even a "moderate" *r* of .3 is a genuinely useful finding — don't dismiss it for not being .9.

A quieter fact about size

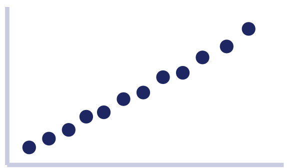
If you **square** *r* you get $\boldsymbol{r^2}$, the share of one variable's spread that "goes along with" the other (we'll meet it again in Unit 7). $r = .50$ means $r^2 = .25$ — about a quarter of the variation tracks together. That's why doubling *r* is much more than doubling the relationship: strength grows by the *square*.

4. Scatterplots — seeing the relationship

A **scatterplot** is the picture behind *r*. It puts **one dot per person**, and each dot is that person's two scores: **x** across the bottom, **y** up the side. The shape of the cloud *is* the correlation — *r* is simply that cloud boiled down to one number.

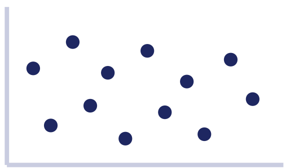
- **Positive** the cloud climbs **uphill** from left to right.
- **Negative** the cloud slides **downhill** from left to right.
- **None** a shapeless **blob** with no tilt; you can't draw an honest trend line through it.

The tighter the dots cling to an imaginary line, the larger |*r*|; the more they spray out, the closer to 0. Here are three clouds, each with the same dozen people, to calibrate your eye:



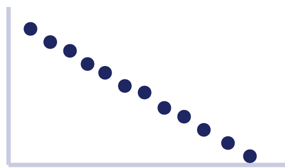
Strong positive (*r* ≈ +.8)

Tight band climbing uphill . High x goes with high y.



Near zero (*r* ≈ 0)

Shapeless blob, no tilt. x tells you nothing about y.



Strong negative (*r* ≈ −.8)

Tight band sliding downhill . High x goes with low y.

Quick test

The picture and the number **always agree**. *r* is just the dot cloud, summarized into one value. If the cloud tilts uphill, *r* is positive; if it tilts downhill, *r* is negative; if it has no tilt, *r* is near 0. So before trusting any reported *r*, picture the cloud it must have come from.

CHECK

On a scatterplot of study hours (x) and GPA (y), the cloud of dots climbs uphill from lower-left to upper-right. The correlation is…

Positive. An uphill cloud means y rises as x rises — more study hours go with higher GPA. That's the definition of a positive relationship, so *r* is positive. A downhill cloud would mean a negative *r*.

5. Worked example — estimating *r* by eye on OUR data

You won't always have a calculator handy — and on exam day you'll want to *sanity-check* whatever number Sheets hands you. So let's reason our way to an estimate from the picture alone. Open the **SJNY Student Life** dataset. We want to know: does **social media** go with **loneliness**? That's `social_media_hours` (column H) and `loneliness` (column L). Plot the cloud, then read it step by step.

Worked example — eyeball the correlation for social media loneliness

Look at the cloud's tilt and tightness, turn that into a sign and rough strength, then interpret it cautiously. Predict each step before you reveal it.

STEP 1 — WHICH WAY DOES THE CLOUD TILT?

Make the scatterplot (x = social media hours, y = loneliness). The cloud climbs **uphill** : people higher on phone time tend to sit higher on loneliness. Uphill means the sign is `positive (+)`.

STEP 2 — HOW TIGHTLY DO THE DOTS HUG A LINE?

The dots cluster in a clear upward band, but with real scatter — not a razor-thin line, not a shapeless blob. A clear-but-fuzzy uphill band reads as a **strong** relationship, somewhere around $r \approx +.5$ on our rough guide.

STEP 3 — CONFIRM WITH THE FORMULA

In any empty cell, type `=CORREL(H2:H201, L2:L201)`. Out comes $r = +.52$ — right where the eyeball estimate said. The picture and the number agree, as they always must.

STEP 4 — INTERPRET, CAUTIOUSLY

Say it in words: "more social media hours go with more loneliness, and the link is strong (r = +.52)." Then **stop** — resist "phones cause loneliness." All we've shown is that the two *travel together*; Section 6 explains why that's not the same as cause.

A few more pairs from our dataset, already computed for you — cover the right column and try to predict each from its description before peeking:

Pair	r	In words
social media × loneliness	+.52	strong, positive
social media × wellbeing	−.46	strong, negative
study hours × GPA	+.48	strong, positive
sleep hours × stress	−.48	strong, negative
animal compassion × self-compassion	+.47	strong, positive
GPA × job hours	−.02	basically **no** relationship

6. The most important caveat — correlation ≠ causation

Here is the single most important — and most ignored — lesson in this whole course: **a correlation cannot prove that one thing causes the other.** Two variables can ride together for reasons that have nothing to do with one steering the other. There are two classic escape hatches, and a correlation alone can't rule either out.

The third-variable problem. Across a summer, **ice-cream sales** and **drownings** rise and fall together — a strong positive correlation. Does ice cream drown people? Of course not. A hidden **third variable** — **hot weather** — drives **both**. More heat means more ice cream *and* more swimming. The two are linked, but neither **causes** the other; a lurking variable pulls both strings at once.

The directionality problem. Even when two things *are* causally linked, a correlation can't tell you which way the arrow points. If phone time and loneliness travel together, maybe phones make people lonely — or maybe lonely people reach for the phone. Same *r* either way; the number is mute on direction.

CHECK **Towns with more ice-cream sales have more drownings (a strong positive correlation). The best explanation is…**

A third variable. Hot weather pushes up both ice-cream buying and swimming (and so drownings) at the same time. The two are correlated, but neither causes the other — the textbook example of why correlation never proves causation.

Now turn the lesson on our own headline finding. We found ***r* = +.52** between phone time and loneliness (a pattern echoed in real research; Primack et al., 2017). But ask three honest questions:

- Does the **phone cause** loneliness?
- Do **lonely people** reach for the phone **more**? (directionality)
- Does something **else** — anxiety, or being new on a commuter campus — drive **both**? (third variable)

The most important lesson

From ***r* alone, you cannot tell.** The number is real; the story behind it is still wide open. Only a true **experiment** — randomly assigning people to conditions (Unit 11 and on) — can pry apart cause from coincidence. A correlation can never tell you what causes what.

7. Two traps that fool *r* — restricted range and outliers

Even when you remember not to claim causation, *r* itself can mislead you about the *strength* of a relationship. Two situations do most of the damage.

Restriction of range. Pearson's *r* measures how scores vary *together*, so it needs real variation to work with. If your sample has been squeezed into a narrow slice — only top students, only people who already use their phones a lot — the correlation can collapse toward zero even when a strong relationship exists in the full population. The classic case: SAT scores barely predict college grades *at a highly selective school*, because everyone admitted already scored high. Chop off the low end and the uphill tilt flattens out. So a small *r* can mean "no relationship" — or it can mean "I only looked at a narrow band." Always ask whether your sample covers the full range.

Outliers. Because *r* is built from means and squared distances, a single extreme point can drag it around dramatically — especially in a small sample. One person who is sky-high on x and y can manufacture a positive correlation out of an otherwise shapeless blob; one oddball in the corner can wipe out a real one. This is exactly why you **plot the data first**: the scatterplot reveals a lone dot living far from the pack that the bare number would never confess. Never report an *r* you haven't *looked* at.

The fix is the same for both

Make the scatterplot before you trust the number. A restricted range shows up as a cramped little cloud; an outlier shows up as a lone dot far from the rest. The picture catches what *r* hides.

And one structural limit worth repeating: *r* only catches **straight-line** patterns. A strongly **curved** relationship — say, arousal and performance, which rises then falls (an upside-down U) — can post an *r* near 0 even though the two are tightly linked. Near-zero *r* means "no *line*," not "no relationship of any kind." Once again: plot first.

The fine print — when Pearson's *r* is trustworthy

Beyond the two traps, *r* makes a few quiet assumptions worth naming. **The relationship is linear** — the structural limit above, promoted to a formal assumption. **Both variables are quantitative** (interval/ratio) — deviations from a mean have to make sense for both. And **the pairs are independent** — one row per person, no duplicated or linked cases. The traps plus these three checks are your full pre-flight list, and one habit screens them all at a glance: *make the scatterplot before you trust the number.*

Is meat male?

Correlation lets us test surprising ideas. Researchers have long suspected a link between **traditional masculinity** and **meat-eating** — that red meat carries a "manly" cultural charge (Rozin et al., 2012). Our dataset lets you check it: `masculinity` against `meat_commitment` gives ***r* = +.35**, a moderate positive relationship. Higher masculinity scores go with stronger commitment to eating meat. It's a real pattern in the data — a moderate-strength uphill cloud — though, as Section 6 warns, a correlation alone can't tell us *why*. Does endorsing masculine norms push someone toward meat, does a meat-heavy upbringing shape how masculine someone feels, or does some third thread — region, family, peer culture — weave through both? *r* = +.35 flags the pattern as worth studying; it doesn't close the case.

In Google Sheets

You won't compute *r* by hand — Sheets does it with one function. Give it two equal-length columns and it returns Pearson's *r*:

```
=CORREL(H2:H201, L2:L201)          r between social media and loneliness  (+.52)
=CORREL(I2:I201, M2:M201)          r between exercise and wellbeing
```

The two ranges **must be the same length** and lined up row-for-row, so that row 2 is the same person in both columns. (`=PEARSON` does the identical job, if you ever see it.) To see the cloud behind the number — which you should, every time — highlight the two columns, then **Insert ▸ Chart**, and pick **Scatter** in the Chart editor. Looking at the picture is how you catch the outliers and curves that *r* quietly hides.

Stat in the Wild — spurious correlations

If you hunt through enough data, you'll find absurd correlations by sheer chance. One famous collection pairs the **U.S. per-capita consumption of cheese** with the **number of people who died tangled in their bedsheets** — *r* above .9 across a decade. Nobody thinks cheese strangles sleepers; the two just happened to drift upward together over the same years.

The lesson: a high *r* is not, by itself, a discovery. With no plausible mechanism and no experiment, even a near-perfect correlation can be pure coincidence — a **spurious** relationship. Demand a reason, not just a number.

Watch out for

A correlation is **not** proof of cause — ever; repeat it like a mantra (third variables and directionality both lurk). The sign is direction, not quality: a negative *r* is a real, strong relationship, not a "bad" one. *r* only catches *straight*-line patterns — a curved relationship can fool it toward 0. A **restricted range** can shrink a real correlation, and a single **outlier** can invent or erase one — so always plot the cloud before you trust the number. And near 0 means no *line* (like GPA × job hours, $r = -.02$); the two just don't track each other.

Key terms

- Correlation
- Correlation coefficient (*r*)
- Pearson's *r*
- Positive relationship
- Negative relationship
- Strength / effect size
- r^2
- Scatterplot
- Third variable
- Directionality
- Correlation ≠ causation
- Restriction of range
- Outlier
- Spurious correlation

Try it yourself — Pearson *r*

Paste two equal-length columns, hit Run, and read a plain-English result. Or load our class data to see the headline $r = +.52$.

30-SECOND GAME

Guess the Correlation

Look at the dot cloud — how tightly does it hug a line? Train your eye and you'll *feel* what ***r*** means before you ever compute it.

PRACTICE — NEVER GRADED

Check yourself

Ten quick true/false items with instant explanations. (Different from your Canvas quiz — real practice.)

Sigma says: When an *r* confuses you, split it in two — read the *sign* for direction, then the *size* for strength. Sign and size are two separate questions, and *r* answers both at once. And no matter how big it is, never let an *r* whisper "cause" in your ear.